МИНОБРНАУКИ РОССИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»

Обработка данных биоэлектрического импеданса клеточных суспензий

Электронные методические указания
к интерактивной практической работе

САМАРА
2011

УДК 57.087

Составители: **Комарова Марина Валерьевна,**
Акулов Сергей Анатольевич,

Рецензент: декан радиотехнического факультета,
к.т.н., доцент Кудрявцев Илья Александрович

Обработка данных биоэлектрического импеданса клеточных суспензий [Электронный ресурс] : электрон. метод. указания к интерактив. практ. работе / Минобрнауки России, Самар. гос. аэрокосм. ун-т им. С. П. Королева (нац. исслед. ун-т); сост. М.В. Комарова, С. А. Акулов. –Электрон. текстовые и граф. дан. (599 Кбайт). – Самара, 2011. – 1 эл. опт. диск (CD-ROM).
Режим доступа: http://rtfmoodle.ssau.ru/course/view.php?id=38

В электронных методических указания приведены основные сведения о методах измерения электрического импеданса клеточных суспензий, описаны основы регрессионного анализа данных о параметрах электрического импеданса клеточных суспензий. Электронные методические указания содержат описание программной оболочки, порядок выполнения практической работы и требования к отчету.

Электронные методические указания предназначены для магистрантов, обучающихся по направлению 201000.68 «Биотехнические системы и технологии» по дисциплине “Биологические системы и технологии” в 9 семестре. Доступ к интерактивной практической работе осуществляется по сетевому адресу:
http://rtfmoodle.ssau.ru/course/view.php?id=38

Разработано на кафедре радиотехники и медицинских диагностических систем.

Цель работы: ознакомиться с проведением регрессионного анализа в среде свободно распространяемого статистического пакета R; построить модели нелинейной регрессии по массивам данных биоэлектрического импеданса клеточных суспензий и данных, полученных в ходе космического эксперимента, проанализировать полученные модели.

1 ТЕОРЕТИЧЕСКИЕ ОСНОВЫ РАБОТЫ

1.1 Регрессионный анализ

Исследование зависимостей — одна из распространённых задач в обработке данных экспериментов и обсервационных исследований биотехнических систем.

Регрессионный анализ — метод установления аналитической зависимости между зависимой переменной y и одной или несколькими независимыми переменными x_1, $x_2,\ldots,x_k$.

Независимые переменные иначе называются *регрессорами* или *предикторами* (от англ. *predict* — предсказывать), а зависимая переменная — *переменной отклика*. Следует отметить, что понятия зависимых и независимых переменных отражает лишь математическую зависимость переменных, но не причинно-следственные отношения.

Регрессия бывает:

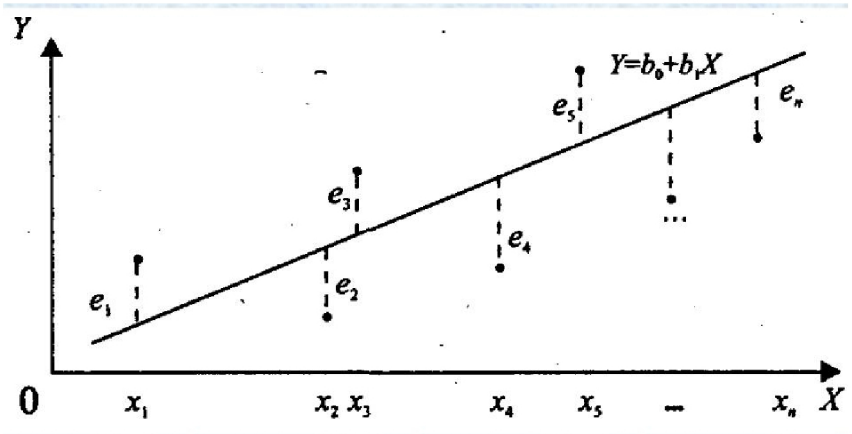
– по числу предикторов — парная и множественная;

– по форме зависимости — линейная и нелинейная.

Рассмотрим вначале *парную линейную регрессию*: $y=b_0+b_1x$. По выборке ограниченного объёма нельзя точно определить теоретические значения параметров (коэффициентов) регрессии, однако можно построить эмпирическое уравнение регрессии. Делается это с помощью *метода наименьших квадратов*, когда минимизируется сумма квадратов отклонений реально наблюдаемых y_i от их оценок $\hat{y}_i$ (имеются в виду оценки с помощью прямой линии, претендующей на то, чтобы представлять искомую регрессионную зависимость): $\hat{y}_i = b_0 + b_1 x_i$.

Пусть по выборке данных (x_i,y_i), $i=1, 2,\ldots, n$, требуется определить оценки b_0 и b_1 эмпирического уравнения регрессии: $y=b_0+b_1x$. Назовём *остатками*, или *ошибками* регрессии разность между наблюдаемым и оцененными значениями переменной отклика

$$e_i = y_i - \hat{y}_i .$$



В методе наименьших квадратов минимизируется функция:

$$Q(b_o, b_b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2. \quad (1)$$

Так как функция $Q(b_0, b_1)$ непрерывна, выпукла и ограничена снизу, то она имеет минимум. Необходимым условием минимума $Q(b_0, b_1)$ является равенство нулю ее частных производных по неизвестным параметрам b_0 и b_1.

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2\sum (y_i - b_0 - b_1 x_i) = 0 \\ \dfrac{\partial Q}{\partial b_1} = -2\sum (y_i - b_0 - b_1 x_i) x_i = 0 \end{cases} \quad (2)$$

$$\begin{cases} \sum y_i - \sum b_0 - \sum b_1 x_i = 0 \\ \sum x_i y_i - \sum b_0 x_i - \sum b_1 x_i^2 = 0 \end{cases}; \quad (3)$$

Разделим на n оба уравнения:

$$\begin{cases} \dfrac{\sum b_0}{n} + \dfrac{\sum b_1 x_i}{n} = \dfrac{\sum y_i}{n} \\ \dfrac{\sum b_0 x_i}{n} + \dfrac{\sum b_1 x_i^2}{n} = \dfrac{\sum x_i y_i}{n} \end{cases}. \quad (4)$$

$$\begin{cases} b_0 + b_b \bar{x} = \bar{y}; \\ b_0 \bar{x} + b_1 \overline{x^2} = \overline{xy}. \end{cases} \quad (5)$$

Решим полученную систему уравнений относительно b_0 и b_1.

$$b_0 = \bar{y} - b_b \bar{x}; \quad (6)$$

$$b_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}. \quad (7)$$

$$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = r\frac{S_y}{S_x}. \quad (8)$$

$$b_0 = \frac{\sum y_i \sum x_i^2 - \sum x_i y_i \sum x_i}{n\sum x_i^2 - (\sum x_i)^2}. \quad (9)$$

Интерпретация коэффициентов регрессии.

Коэффициент b_1 равен тангенсу угла наклона между прямой регрессии и осью абсцисс и показывает, как изменяется переменная отклика при увеличении предиктора на единицу.

Коэффициент b_0, или свободный член равен значению переменной отклика при нулевом значении независимой переменной.

Множественная линейная регрессия позволяет изучить совместное воздействие нескольких независимых переменных на переменную отклика.

$$y=b_0+b_1x_1+b_2x_2 \ldots +b_kx_k$$

Коэффициенты множественной линейной регрессии, как и в случае парной зависимости рассчитывают также с помощью метода наименьших квадратов, но в матричной записи. Интерпретация коэффициентов регрессии та же, что и в парной регрессии — на сколько изменится переменная отклика при увеличении предиктора на единицу.

*Нелинейные регрессионные модели***.** Многие связи по своей природе не линейны, это следует из физических, химических, биологических, социально-экономических и других предметных закономерностей. Для оценки параметров нелинейных моделей используют два подхода.

В одних случаях возможна линеаризация модели путем преобразования исходных переменных. Такие модели оказываются линейными по параметрам, хоть они и нелинейны по переменным.

Примеры:

$y = b_0+b_1x+b_2x^2$ — параболическая регрессия;

$y = b_0+b_1x+b_2x^2+b_3x^3$ — полином 3-йстепени;

$y = b_0+b_1x^{-1}$ — гиперболическая регрессия.

Если модель нелинейна по переменным, то введением новых переменных её можно свести к линейной модели. Например, в полиномиальной регрессии можно ввести новые переменные: $x_2=x^2$, $x_3=x^3$. Для оценки параметров в таком случае можно использовать метод наименьших квадратов.

Совершенно иначе обстоит дело, когда линеризующее преобразование подобрать не удаётся, например, зависимость представлена степенной функцией, в которой показатель степени является параметром. В таких случаях говорят, что регрессия нелинейна по оцениваемым параметрам, и именно такие модели в регрессионном анализе считаются действительно нелинейными.

Примеры некоторых часто употребляемых нелинейных функций.

1) Асимптотические функции:

$$y = \frac{ax}{1+bx};$$

$y = a(1-e^{-bx})$ — двухпараметрическая асимптотическая экспоненциальная;

$y = a-be^{-cx}$ — трёхпараметрическая асимптотическая экспоненциальная;

2) S-образные функции:

$y = \frac{e^{a+bx}}{1+e^{a+bx}}$ — двухпараметрическая логистическая;

$y = \frac{a}{1+be^{-cx}}$ — трёхпараметрическая логистическая;

$y = a + \frac{b-a}{1+e^{(c-x)/d}}$ — четырёхпараметрическая логистическая;

$y = a - be^{-(cx^d)}$ — Вейбулла;

$y = ae^{-be^{-cx}}$ — Гомперца;

3) «Горбатые» кривые:

$y = axe^{-bx}$;

$y = ae^{(-|bx|^2)}$;

$y = ae^{bx} - ce^{-dx}$.

Нахождение параметров нелинейной регрессии как и в случае линейных моделей основано на методе наименьших квадратов, но имеет свои особенности. Запишем в общем виде нелинейную регрессионную модель:

$$y=f(\beta, x_i),$$

где β — вектор параметров, а x_i — вектор предикторов. Необходимо минимизировать сумму квадратов отклонений эмпирических значений признака от теоретических, полученных по уравнению регрессии:

$$Q(\beta) = \sum_{i=1}^{n} (y_i - f(\beta, x_i))^2 . \quad (10)$$

Дифференцируя функцию $Q(\beta)$ получаем:

$$\frac{\partial Q(\beta)}{\partial \beta} = -2\sum_{i=1}^{n} (y_i - f(\beta, x_i)) \frac{\partial f(\beta, x_i)}{\partial \beta} . \quad (11)$$

Приравнивая частные производные по каждому из параметров к нулю, получаем систему уравнений для нахождения коэффициентов регрессии. Однако, из-за нелинейности моделей применяют численные методы дифференцирования, которые реализуются в компьютерных программах итеративным путём.

На практике в различных прикладных статистических пакетах пользователю предлагается задать форму зависимости и вести стартовые (исходные) значения параметров. Дальнейшая работа вычислительного алгоритма следующая. Вычисляются предсказанные значения y по фактическим значениям x с использованием заданных значений параметров регрессии. Вычисляются остатки для всех наблюдений в выборке и затем сумма квадратов остатков. Вносятся небольшие изменения в одну или более оценку параметров. Вычисляются новые предсказанные значения y, остатки и сумма квадратов остатков. Если сумма квадратов остатков меньше, чем прежде, то новые оценки

параметров лучше прежних и их следует использовать в качестве новой отправной точки. Последние три шага повторяются вновь до тех пор, пока не окажется невозможным внести такие изменения в оценки параметров, которые привели бы к изменению суммы остатков квадратов.

Оценка качества уравнения регрессии

Цель проверки качества уравнения регрессии *в целом* — оценить насколько хорошо эмпирическое уравнение регрессии согласуется со статистическими данными. На практике выполняют как проверку качества модели в целом, так и отдельных параметров регрессии. Основные показатели качества регрессионной модели:

1. Значение F-статистики (применяется в линейной регрессии).
2. Коэффициент детерминации R^2.
3. Стандартная ошибка регрессии.
4. Стандартная ошибка параметра регрессии.
5. Статистическая значимость параметра регрессии

Рассмотрим перечисленные показатели качества более подробно.

1. Если хотя бы один коэффициент статистически значим (т.е. отличен от нуля), то говорят, что регрессия существует. Проверку существования регрессии начинают с проверки значимости уравнения в целом с помощью дисперсионного анализа. В основе лежит разделение полной суммы квадратов отклонений на две части: объясненные моделью и из-за ошибок:

$$SS_{tot} = SS_{reg} + SS_{res}\,, \tag{12}$$

где:

$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ — полная сумма квадратов отклонений (англ. *total sum of squares*);

$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ — сумма квадратов отклонений, объяснённая моделью (англ. *regression sum of squares*);

$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y})^2$ — остаточная сумма квадратов отклонений (англ. *residual sum of squares*).

Из этих величин конструируется F-статистика:

$$F = \frac{SS_{reg} / df_1}{SS_{res} / df_2}\,, \tag{13}$$

где:

df — число степеней свободы (англ. *degree of freedom*);

$df_1 = k$ — число независимых предикторов,

$df_2 = n–k–1$.

F-статистика имеет распределение Фишера с числом степеней свободы df_1 и df_2. Если вычисленное значение F-статистики больше критического значения (определяют по таблицам) на заданном уровне значимости α (1%, 5%, 10%), то уравнение регрессии в целом является статистически значимым на заданном уровне значимости. В противном случае построенная модель в целом не значима. В современных компьютерных

статистических пакетах выводится как значение F-статистики, так достигнутый уровень значимости.

2. Коэффициент детерминации модели:

$$R^2 = \frac{SS_{reg}}{SS_{total}} = 1 - \frac{SS_{res}}{SS_{tot}}. \quad (14)$$

Коэффициент детерминации показывает, какая доля дисперсии переменой отклика объясняется построенной моделью. Он отражает силу влияния на переменную отклика нескольких независимых переменных. Например, при R^2 равным 0,65 регрессионная модель объясняет 65% дисперсии переменой отклика, остальные же 35% вариации объясняются прочими, неучтёнными факторами.

Коэффициент детерминации достигает максимума из возможных значений, когда остаточная сумма квадратов отклонений принимает наименьшее из достижимых значений. Критерий максимума R^2 эквивалентен принципу метода наименьших квадратов.

Коэффициент детерминации изменяется в диапазоне от 0 до 1. Чем ближе R^2 к 1, тем лучше регрессия аппроксимирует статистические данные, тем теснее связь между зависимой и объясняющими переменными. Если $R^2 = 1$, то статистические данные лежат на линии регрессии, т.е. между зависимой переменной и предикторами существует функциональная зависимость. Если же $R^2 = 0$, то вариация переменной отклика полностью обусловлена воздействием не учтённых в модели факторов.

На практике R^2 часто используют для предварительной оценки качества регрессионной модели. Только модели с высоким коэффициентом детерминации пригодны для задач предсказания зависимой переменной по известным предикторам. В то же время, в ряде прикладных задач, например, по анализу экологических и биомедицинских систем представляют интерес и статистически значимые модели с невысоким R^2, позволяющие выявить набор предикторов из числа потенциально возможных, которые оказывают наибольшее влияние на отклик.

У коэффициента детерминации есть ряд недостатков. R^2 всегда увеличивается с включением новой переменной в модель. Коэффициента детерминации в разных моделях с разным числом переменных и наблюдений несравнимы. Следует также иметь в виду, что R^2 является смещенной оценкой. Корректированная оценка коэффициента детерминации получается по формуле:

$$R^2_{correct} = 1 - \frac{SS_{res}/(n-k-1)}{SS_{tot}/(n-1)}. \quad (15)$$

Величину $R = \sqrt{R^2}$ называют множественным коэффициентом корреляции. Он отражает взаимосвязь между между наблюдаемыми и расчетными значениями y. В случае парной регрессии коэффициент корреляции между переменными x и y равен по абсолютному значению множественному коэффициенту корреляции, и соответственно коэффициент детерминации модели квадрату коэффициента корреляции.

После того, как определён коэффициент детерминации, можно по-другому представить F-статистику. Выразим сумму квадратов отклонений регрессии и остатков через общую сумму квадратов отклонений и коэффициент детерминации и подставим в формулы для F-статистики.

$$SS_{res} = SS_{tot} - SS_{reg} = SS_{tot} - R^2 \times SS_{tot} = SS_{tot}(1 - R^2);$$

$$SS_{res} = SS_{tot} \times R^2;$$

$$F = \frac{SS_{reg}/df_1}{SS_{res}/df_2} = \frac{R^2 \cdot SS_{tot}/df_1}{(1-R^2)\cdot SS_{tot}/df_2} = \frac{R^2/df_1}{(1-R^2)/df_2}. \quad (16)$$

Подобное представление F-статистики применяют в пошаговых алгоритмах отбора предикторов в множественных линейных моделях, когда оценивают, насколько изменится F-статистика при добавлении предиктора в модель (или, наоборот, исключении предиктора из модели). Если изменение F-статистики с введением нового предиктора в уравнение множественной регрессии больше некоторого критического уровня, то такой предиктор целесообразно оставить в модели.

3. Стандартная ошибка регрессии. О качестве полученного уравнения регрессии можно судить, исследовав оценки случайных ошибок уравнения. Оценка дисперсии случайной ошибки рассчитывается по формуле

$$S^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-k-1}. \quad (17)$$

Величина S называется стандартной ошибкой регрессии. Чем меньше величина S, тем лучше уравнение регрессии описывает независимую переменную y.

4. В регрессионном анализе можно оценить не только качество модели в целом, но и отдельных параметров модели. Для этого рассчитывают стандартные ошибки коэффициентов регрессии — $SE(b_i)$:

$$SE(b_i) = \sqrt{\frac{S^2}{\sum (x_i - \bar{x})^2}}. \quad (18)$$

5. С помощью t-статистики проверяют нулевую гипотезу H_0: $b_i = 0$ против альтернативной гипотезы H_1: $b_i \neq 0$:

$$t = \frac{b_i}{SE(b_i)}. \quad (19)$$

По величине t, имеющей распределение Стьюдента, компьютерные статистические пакеты вычисляют p-значение. Если p меньше заданного уровня значимости, то нулевую гипотезу отвергают на уровне значимости α. В этом случае можно говорить о статистической значимости данного параметра регрессии.

6. Для получения информации об адекватности уравнения регрессии исследуют регрессионные остатки. Если выбранная модель хорошо описывает истинную зависимость, то остатки должны быть независимыми нормально распределенными

случайными величинами с нулевым средним, и в их значениях должен отсутствовать тренд. Анализ остатков регрессии — это процесс проверки выполнения этих условий. Применяют графические и аналитические методы анализа остатков.

– Диаграмма рассеивания: предсказанные значения (ось абсцисс) — остатки или нормированные остатки, т.е. делённые на дисперсию остатков (ось ординат). Наличие криволинейного тренда — плохо. Случайный разброс — хорошо.

– График нормированных остатков в зависимости от номера объекта. Позволяет заметить тренд от времени.

– Расстояния Кука — мера влияния соответствующего наблюдения на уравнение регрессии. Эта величина показывает разницу между вычисленными *b*-коэффициентами и значениями, которые получились бы при исключении соответствующего наблюдения. В адекватной модели все расстояния Кука должны быть примерно одинаковыми; если это не так, то имеются основания считать, что соответствующее наблюдение (или наблюдения) смещает оценки коэффициентов регрессии.

– Статистика Дарбина-Уотсона применяется для проверки независимости остатков друг от друга:

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}, \qquad (20)$$

где $e_i = y_i - \hat{y}_i$.

Статистика D изменяется от 0 до 4 и интерпретируется следующим образом:

D=2 — нет автокорреляции остатков

D<<2 — положительная серийная корреляция, то есть близки друг к другу (меньше 1 — признак тревоги)

D>>2 — отрицательная серийная автокорреляция, возможна неодооценка уровня значимости регрессии.

1.2 АНАЛИЗ ДАННЫХ В СРЕДЕ ПАКЕТА R

R представляет собой набор программных средств для управления данными, вычислений и графического отображения. Пакет R имеет модульную структуру, в стандартный комплект входит около 25 модулей, сотни других могут быть поставлены дополнительно, исходя из решаемых задач. Официальный сайт пакета R: http://cran.r-project.org. На многочисленных зеркалах R можно скачать виде исходных кодов или дистрибутивов, скомпилированных под различные операционные системы (Linux, MacOS, Windows).

В числе возможностей пакета R:

* эффективная обработка и хранение данных;
* набор операторов для обработки массивов, в частности матриц;
* цельная, непротиворечивая, комплексная коллекция утилит для анализа данных,
* графические средства для анализа данных и визуализации либо непосредственно на компьютере или при выводе на печать, а также

* хорошо развитой, простой и эффективный язык программирования, включающий условия, циклы, определенные пользователем рекурсивные функции и возможности ввода–вывода.

Термин "окружение"/"среда" введён, что бы подчеркнуть наличие полностью спланированной и непротиворечивой системы, а не постепенно возникшего конгломерата специфических и негибких утилит, как это бывает с другим программным обеспечением анализа данных. R выступает как средство разработки новых методов интерактивного анализа данных. Он динамично разрабатывается, и постоянно пополняется большой коллекцией пакетов. Необходимо отметить, что R — не только статистическая система. Правильнее считать R окружением, в котором реализованы многие классические и современные статистические методы.

В отличие от других программных комплексов, предназначенных для статистического анализа данных, таких например, как SAS, SPS, Statistica, R выводит минимум результатов и сохранит вывод в подогнанный объект для последующего использования в дальнейшем в вызываемых R функциях.

При использовании R программу она выводит приглашение в момент ожидания ввода команд. По умолчанию приглашение это ">", и может показаться, что ничего не произошло.

Получение помощи по функциям и возможностям

R имеет встроенную команду help. Чтобы получить дополнительную информацию о какой-либо конкретной функции, например solve, нужно ввести:

```
> help(solve)
```

Альтернативный вариант

```
>?solve
```

Для наименования обозначенного специальными символами, аргумент должен быть помещен в двойные или одиночные кавычки, что превращает его в "строку символов": Это также необходимо для некоторых слов имеющих синтаксическое значение, включая if, for и функции.

```
> help("[[")
```

Любая форма кавычек может быть использована, чтобы экранировать другие кавычки, как в строке "It's important".

Команда help.search позволяет искать подсказку различными способами: выполните ?help.search для подробностей и конкретных примеров. Примеры по теме страницы помощи можно, как правило, запустить командой:

```
> example(тема)
```

Windows версии R имеют другие дополнительные системы помощи:

```
>?help
```

Команды R и учет регистра.

Технически R является языком выражений с очень простым синтаксисом. Будучи изначально созданным под UNIX, он учитывает регистр. Так "A" и "a" различные символы и будут обозначать различные переменные. Набор символов, которые могут быть

использованы как имена в R зависит от операционной системы и страны, в которых будет запущен R. Обычно разрешены все алфавитно-цифровые символы (а в некоторых странах включая и акцентированные буквы) вместе с "." и "_", с ограничением, что имя должно начинаться с "." или буквы, а если оно начинается с "." второй символ не должен быть цифрой.

Простые команды состоят из выражений либо присвоений.

– Если выражение вводится как команда, оно вычисляется, выводится (если специально не сделано невидимым), и результат теряется.

– Присвоение также вычисляет выражение и передает значение переменной, но результат автоматически не выводится.

Команды разделяются либо точкой с запятой (";"), или переводом строки. Простые команды могут быть сгруппированы в единое составное выражение фигурными скобками ("{" и "}").

Комментарии могут быть практически где угодно, начинаются с символа решетки ("#"), при этом все до конца строки является комментарием.

Если команда не завершена в конце строки, R выдаст особое приглашение, по умолчанию + во второй и последующих строках и продолжит ожидать ввода пока команда не будет синтаксически завершена. Это приглашение может быть изменено пользователем.

Повтор и коррекция предыдущих команд.

Во многих версий UNIX и Windows, R предусматривает механизм восстановления и повторного выполнения предыдущей команды. Вертикальные стрелки на клавиатуре можно использовать для прокрутки вперед и назад по истории команд. Когда команда найдена таким способом, курсор может перемещаться внутри команды с помощью горизонтальных стрелок, можно удалять символы при помощи клавиши <DEL> или добавлять при помощи остальных клавиш.

Сохранение данных и удаление объектов

Записи, которые R создает и которыми манипулирует, называются объекты. Это могут быть переменные, массивы чисел, строки символов, функции, или более сложные структуры, построенные из этих компонентов. В ходе сессии R создаются и хранятся поименованные объекты. R команда

```
> objects()
```

может быть использована для отображения названий объектов, которые в настоящее время хранятся в R. Набор объектов который в настоящее время хранится называется рабочее пространство. Чтобы удалить объекты доступна функция rm:

```
> rm(x, y, z, ink, junk, temp, foo, bar)
```

Все объекты, созданные в ходе R сессий могут быть сохранены в файл для использования в последующих R сессиях. В конце каждой сессии R предоставляется возможность сохранить все имеющиеся в настоящее время объекты. При подтверждении объекты записываются в файл .RData в текущем каталоге, а строки команд, использованных в сессии сохраняются в файл .Rhistory.

Вектора и присваивания

R оперирует именованными структурами данных. Простейшая такая структура это численный вектор, который представляет собой совокупность, состоящую из упорядоченного набора чисел. Чтобы создать вектор с именем x, например состоящий из пяти чисел, а именно 10.4, 5.6, 3.1, 6.4 и 21.7, используют такую команду R:

```
> x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
```

Это выражение присваивания, использует функцию c(), которая в этом контексте может содержать произвольное число аргументов векторов и значение которой есть вектор полученный путем объединения аргументов край в край Одиночное число входящее в выражение трактуется как вектор единичной длины. Учтите, что оператор присваивания ("<-"), который состоит из двух символов "<" ("меньше") и "-" ("минус") выполняется строго однонаправлено и "указывает" на объект получающий значение выражения. В большинстве случаев в качестве альтернативы может быть использован оператор "=". Присваивание можно также сделать с помощью функции assign(). Эквивалентный способ присвоения, приведенному выше, выглядит:

```
assign("x", c(10.4, 5.6, 3.1, 6.4, 21.7))
```

Если выражение используется как законченная команда, его значение выводится и теряется. Так, если мы теперь введем команду:

```
> 1/x
```

то частное пяти элементов будет выведено на терминал (и значение x, разумеется, не изменится). Дальнейшее присваивание

```
> y <- c(x, 0, x)
```

создаст вектор y с 11 элементами, состоящий из двух копий x с нулем между ними.

Действия над векторами

Вектора могут быть использованы в арифметических выражениях, и в этом случае операции выполняются элемент за элементом. Вектора, включающиеся в одно и то же выражение, не обязательно должны быть одной длины. Если длины отличаются, результат выражения это вектор с такой же длиной, как самый длинный вектор, который встречается в выражении. Короткие векторы в выражении используются повторно столько раз, сколько это необходимо (возможно не целое число раз), до тех пор, пока они не совпадут с длинной самого длинного вектора. В частности константа будет просто повторяться. Так, учитывая предыдущие присваивания, команда

```
> v <- 2*x + y + 1
```

создает новый вектор v длинной 11 составленный путем сложения, элемент с элементом, 2*x повторенного 2,2 раза, y повторенного только раз, и 1 повторенной 11 раз.

Элементарные арифметические операторы: +, -, *, / и ^ для возведения в степень. Дополнительно присутствуют функции: log, exp, sin, cos, tan, sqrt, и т.д., все имеют обычное значение.

Функции max и min выделяют соответственно наибольший и наименьший из элементов вектора.

Функция range — это функция, значение которой вектор из двух элементов, а именно c(min(x), max(x)).

Функция length(x) возвращает количество элементов в x.

Функция sum(x) возвращает сумму значений элементов в x, а prod(x) их произведение.

Две статистические функции это mean(x), которая вычисляет выборочное среднее и var(x) — выборочную дисперсию.

2 ПОРЯДОК ВЫПОЛНЕНИЯ РАБОТЫ

Для выполнения данной работы необходимо получить у преподавателя файл с исходными данными. Файл данных имеет текстовый формат и состоит из двух колонок чисел. В первом столбце представлено время (мс), во втором результаты измерений биоэлектрического импеданса клеточных суспензий. Первая строка файла данных содержит имена столбцов (имена переменных в терминологии статпакетов). Фрагмент файла исходных данных:

```
time  imp21
1     0.00E+00
2     9.26E-03
3     2.01E-02
4     3.25E-02
5     4.53E-02
6     6.02E-02
7     7.68E-02
8     9.50E-02
9     1.15E-01
10    1.36E-01
11    1.59E-01
12    1.84E-01
  ...
```

Ниже приведён порядок работы на примере одного из массивов данных, набор команд R, которые необходимо ввести в командной строке, интерпретация выдаваемых сообщений. Команды, вводимые пользователем, в настоящем пособии представлены таким шрифтом. Результаты работы пакета R представлены `таким шрифтом`.

1. Открываем файл данных 21new_.txt и сохраняем его в массиве d21 (data frame в терминологии R). Для этого используем функцию read.table, первым аргументом служит путь до файла. Имя фрейма данных выбирает пользователь на своё усмотрение.

```
d21 <-read.table ("d:/docum/sta_user/mk_ssau_2011/21new_.txt", header=T)
attach (d21)
```

Выведем на экран список имён переменных фрейма данных.

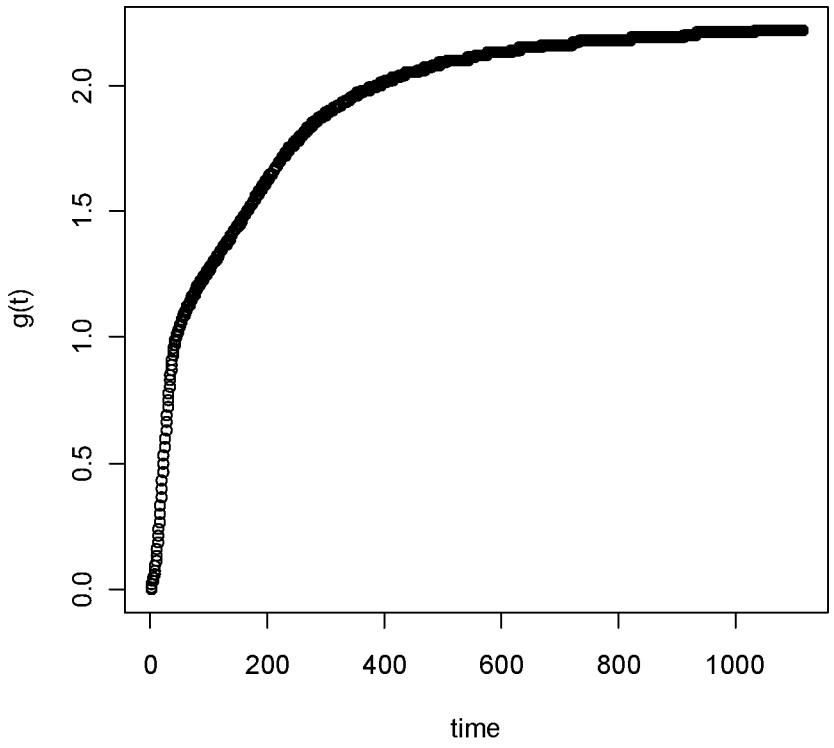
```
names (d21)
```

В ответ на последнюю команду должен появиться список имён исследуемых переменных.

```
[1] "time"  "imp21"
```

2 Строим и анализируем парный график:

```
plot (imp21~time, ylab="g(t)" )
```



Видно, что полученная зависимость нелинейна, можно попробовать применить двух- или трёхпараметрическую экспонециальную аппроксимацию.

3. Построим модель нелинейной регрессии с *двумя* параметрами. Данная зависимость для задания формулы статистическому пакету будет иметь вид:

$$y = a\,(1 - e^{-cx}),$$

При моделировании проведения тока живыми клетками с помощью эквивалентных схем данной зависимости соответствует параллельная схема замещения (рис.1). Переходная функция импеданса: $U = IR(1 - e^{-t/RC})$.Таким образом, IR = a; RC = 1/c.

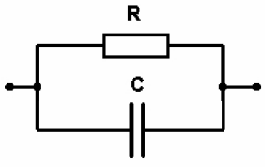


Рис. 1 — Параллельная схема замещения

Для вычисления коэффициентов регрессии применяем функцию nls. В её аргументы входит формула зависимости и начальные (стартовые) значения параметров *a*, *c*.

Вначале пользователю необходимо ориентировочно задать параметры регрессии, которые в процессе итеративного подбора будут заменены на «правильные». Для этого воспользуемся парным графиком разброса. Так, на приведённом рисунке видно, что

кривая асимптотически приближается к значению чуть больше 2, значит параметру a можно присвоить стартовое значение 2. Параметр c можно приблизительно оценить по какой-нибудь точке на графике, подставив её координаты в уравнение с двумя другими уже заданными параметрами. Так, например, рассмотрим точку (200; 1,6). Составим уравнение $1{,}6 = 2 - 2\,e^{-200c}$, откуда $c \approx 0{,}01$.

Результаты нелинейного оценивания сохраним в объект, назовём его fit1 (от англ. *fit* — подгонка).

```
fit1 <- nls (imp21 ~ a *(1-exp(-c * time)),  start = list (a=2,  c= 0.01))
```

Пояснения

- fit1– имя объекта, в который функция nls возвращает результаты работы.
- Оператор "~" используется для определения формулы модели в R. imp21– имя переменной отклика, time — имя независимой переменной, a, в, c — параметры регрессии (имена параметрам можно присвоить произвольно).
- В ответ на данную команду на экране ничего произойти не должно.
- Возможные ошибки в ответ на ввод данной команды:

```
А) Ошибка в nlsModel(formula, mf, start, wts) :
  сингулярная градиентная матрица в оценке начальных параметров
Б) Ошибка в numericDeriv(form[[3L]], names(ind), env) :
  Пропущенное значение или неопределенность получено при вычислении
модели
В) Ошибка в nls(imp21 ~ a - d * exp(c * time + b), data = d22, start =
list(a = 3,  :
  параметры без стартового значения в 'data': imp21, time
```

Пути устранения ошибок: проверить правильность написания команд и имён переменных; изменить начальные значения параметров модели (возможно на порядок!), изменить вид модели.

С помощью функции summary выведем на экран параметры полученной модели.

```
summary (fit1)
```

```
Formula: imp21 ~ a * (1 - exp(-c * time))

Parameters:
   Estimate Std. Error t value Pr(>|t|)
a 2.161e+00  3.346e-03   645.8   <2e-16 ***
c 7.932e-03  6.651e-05   119.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.08476 on 1111 degrees of freedom
```

Трактовка результатов

- `Estimate` — параметр регрессии;
- `Std. Error` — ошибка параметра регрессии;
- `t value` — статистика *t* для оценки уровня значимости параметра регрессии (`t = Estimate/ Std. Error);`
- `Pr(>|t|)` — достигнутый уровень значимости коэффициента регрессии. Здесь тестируется нулевая гипотеза о равенстве нулю параметра регрессии; рассчитанная статистика *t* имеет распределение Стьюдента;
- Signif. codes — выделение звёздочками статистически значимых параметров.
- Residual standard error — стандартная ошибка регрессии.

Таким образом, получено следующее уравнение исследуемой зависимости:

$$imp21 = 2{,}16(1 - e^{-0{,}00793time}).$$

Из таблицы результатов видно, что оба параметра статистически значимые ($p<0{,}001$). Стандартная ошибка регрессии равна 0,085.

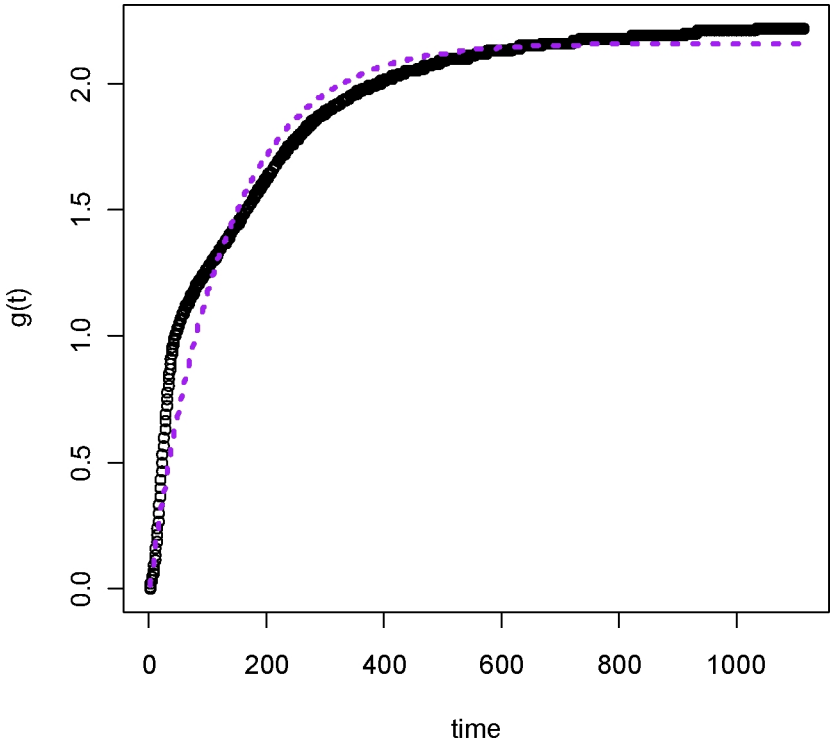
4. Подсчитаем предсказанные по модели значения и добавим их на график разброса.

Для расчёта предсказанных значений воспользуемся функцией predict. Её аргументы: имя модели и вектор данных; мы воспользуемся имеющейся переменной time.

```
pred1 <- predict (fit1, time)
```

Для добавления линии аналитической зависимости на график используем функцию lines. Её аргументы: вектора данных, которые мы хотим добавить на график и соединить линиями. Дополнительные, необязательные аргументы — цвет (col), тип линии (lty), толщина линии (lwd) и др.

```
lines (pred1 ~ time, col= "purple", lty="dotted", lwd=3)
```



5. Построим модель нелинейной регрессии с *тремя* параметрами. Данная зависимость для задания формулы статистическому пакету будет иметь вид:

$$y = a - b\,e^{-cx}.$$

При моделировании проведения тока живыми клетками с помощью эквивалентных схем данной зависимости соответствует последовательно-параллельная схема замещения (рис.2). Переходная функция импеданса: $U = R_0 I + RI\ (1 - e^{-t/RC})$.Таким образом, при единичном токе R_0 +R = a; R=b; RC = 1/c.

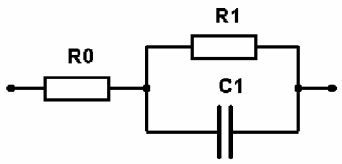


Рис. 2 — Последовательно-параллельная схема замещения

Как и в случае двухпараметрической модели для вычисления коэффициентов регрессии применяем функцию nls. В её аргументы входит формула зависимости и стартовые значения параметров *a*, *b*, *c*. Для подбора стартовых значений параметров воспользуемся парным графиком разброса и оцененными нами ранее параметрами $a \approx 2$ и $c \approx 0{,}01$. В точке (0; 0) параметры *a* и *b* равны, значит $b \approx 2$.

```
fit2 <- nls (imp21 ~ a- b*exp(-c* time),  start = list (a=2, b= 2, c= 0.01))
```

С помощью функции summary выведем на экран параметры полученной модели.

```
summary (fit2)
```

```
Formula: imp21 ~ a - b * exp(-c * time)

Parameters:
   Estimate Std. Error t value Pr(>|t|)
a 2.190e+00  2.756e-03   794.7   <2e-16 ***
b 1.827e+00  9.530e-03   191.8   <2e-16 ***
c 6.189e-03  5.858e-05   105.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06019 on 1110 degrees of freedom
```

Трактовка результатов

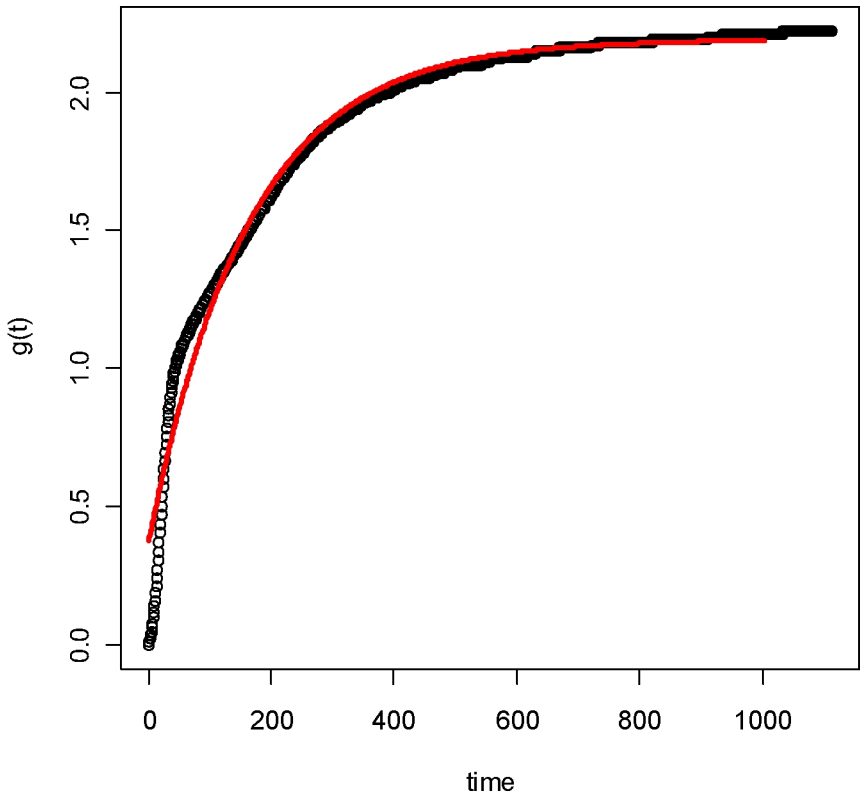
Получено следующее уравнение исследуемой кривой:

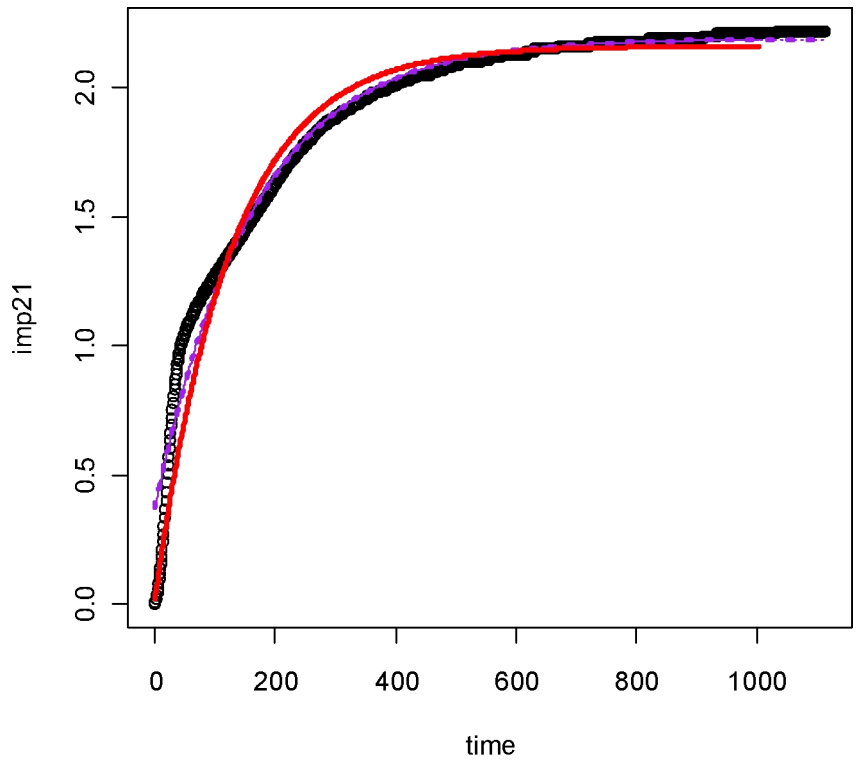
$$imp21 = 2{,}19 - 1{,}827e^{-0{,}00619time}.$$

Из таблицы результатов видно, что все три параметра регрессии (a, b, c) статистически значимые (p<0,001). Стандартная ошибка регрессии равна 0,060.

5. Подсчитаем предсказанные по модели значения и добавим их на график разброса.

```
pred2 <-predict (fit2, time)
lines (pred2 ~ time, col= "red", lwd=3)
```





6. С помощью функции anova сравним построенные модели fit1 и fit2 друг с другом.

anova (fit1, fit2)

```
Analysis of Variance Table

Model 1: imp21 ~ a * (1 - exp(-c * time))
Model 2: imp21 ~ a - b * exp(-c * time)
  Res.Df Res.Sum Sq Df Sum Sq F value    Pr(>F)
1   1111     7.9825
2   1110     4.0215  1  3.961  1093.3 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Большое значение F критерия (F=1093,3) и соответственно низкое значение вероятности нулевой гипотезы (Pr(>F) свидетельствует о различии построенных моделей. Какую модель лучше выбрать для дальнейшей работы? Можно остановить свой выбор на модели 1, как более простой (с меньшим числом параметров), а можно на модели 2, как имеющей меньшую стандартную ошибку регрессии. Обоснуйте свой выбор.

3 СОДЕРЖАНИЕ ОТЧЕТА

1. Цель работы.
2. График разброса исходных данных
3. Параметры моделей нелинейной регрессии и оценка качества регрессии.
4. Графики полученных аналитических зависимостей, наложенные на графики разброса.
5. Вывод о полученных результатах.